DOCUMENT RESUME

ED 326 594 UD 027 815

AUTHOR Murray, Steve
TITLE Commentary on the Gap Reduction Model for Chapter 1 Evaluation. Draft.
INSTITUTION Northwest Regional Educational Lab., Portland, Oreg.
REPORT NO TAC-B-65
PUB DATE 6 Jun 88
NOTE 14p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Compensatory Education; Correlation; Educational Assessment; Elementary Secondary Education; Equated Scores; *Evaluation Methods; *Outcomes of Education; *Program Evaluation; School Districts; *Special Programs; State Departments of Education; Validity
IDENTIFIERS *Education Consolidation Improvement Act Chapter 1; *Title I Evaluation and Reporting System

ABSTRACT

The gap-reduction model has been identified as a potential alternative to or extension of the Title I Evaluation and Reporting System (TIERS). The gap-reduction model has been recommended for the evaluation of bilingual programs, but has only recently been given consideration for evaluating local Chapter 1 programs. This report recommends that projects use local comparison groups and examine multiple outcomes, but voices concern that state educational agencies (SEAs) and local educational agencies (LEAs) will lose sight of the benefits of the gap-reduction model because of significant technical problems with interpreting Relative Growth Indices (RGIs) as an extension of routine Model A analysis. The report discusses the following three issues: (1) the validity of the RGI; (2) conditions for aggregation; and (3) data quality. The following recommendations are presented: (1) do not require aggregating RGIs, but rather encourage piloting the gap-reduction model as a means to augment local evaluations; and (2) encourage LEAs to use annual testing and explicitly allow selection on the pretest at the local level, then leave the gains "uncorrected," acknowledging that they are measures of relative gain, or adjust the results downward to better estimate treatment effects. The report includes one figure and four sample tables. (AF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

TNC-R-65

SCOPE OF INTEREST NOTICE
The ERIC Facility has assigned this document for processing to: UD
In our judgment, this document is also of interest to the Clearinghouses noted to the right. Indexing should reflect their special points of view. TM

ED326594

DRAFT

Commentary on the Gap Reduction Model for Chapter 1 Evaluation

Steve Murray
Region 4 Chapter 1 Technical Assistance Center
Northwest Regional Educational Laboratory

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

June 6, 1988

ED's national evaluation standards issue paper prepared for the regional meetings identifies the gap-reduction model as a potential alternative to, or extension of TIERS. The gap-reduction model has been recommended for evaluating Bilingual programs and it has been disseminated nationally for that purpose, but it has only recently been given consideration for evaluating local Chapter 1 programs. The regional meetings, while one forum for considering the gap-reduction model, are a less than ideal context for such an assessment. Many of those attending the regional meetings will have little awareness of the gap-reduction model and cannot hope to give an informed opinion about its use for Chapter 1 evaluation. Their concerns, however, will focus upon the interpretation of results and implementation demands.

We strongly support encouraging projects to use local comparison groups, which the gap-reduction concept does. Stringfield and Davis (1987) endorsed use of local comparison groups (e.g., schoolwide trends, previous years Chapter 1 evaluations, matched groups of students) to augment local Chapter 1 evaluations (enclosed). Many others have advocated similar views over the years. TIERS itself includes a comparison group model (Model B) for local evaluation.

We also support encouraging projects to examine multiple outcomes, including nontest outcomes, and are grateful to have a simple evaluation model to examine such outcomes. We are concerned, however, that SEAs and LEAs will lose sight of these benefits of the gap-reduction model because of significant technical problems with interpreting Relative Growth Indices as an extension of routine Model A analyses.

UD 027 815

We raise the following three more specific concerns:

1. Although the gap-reduction model is intuitively appealing, the Relative Growth Index, which we assume is being proposed for aggregation, is not well documented. Evidence from previous studies suggests that RGIs are unstable often yielding results that are inconsistent with other analyses of the same data. We question the validity of the RGI.

2. The conditions under which RGIs could be meaningfully aggregated are not specified in the Bilingual Education Evaluation System documentation nor are they self evident. We question the conditions for aggregation.

3. Requiring gap-reduction calculations will increase data processing demands on local districts which would lead to resistance and considerable quality control problems. We question the impact on data quality.

We discuss each of these three points in order.

Validity of the RGI

Our most fundamental concern with requiring the gap-reduction calculations (Relative Growth Indices) is that RGIs are predictably unstable. In some instances the results, even when calculated strictly according to the recommended procedures, will be uninterpretable. Gabriel (1982) speaks directly to this point (enclosure). His analysis responded to an earlier ED request to examine the conversion of NCE gains to a more understandable metric, which at that time was called percent additional growth but which is nearly identical to the RGI. Gabriel examined interpretation problems with percent additional growth indices that were traced to the standard scores and the denominator in the formula. This denominator is the standardized growth for a comparison group (e.g. national norm group) not receiving program services.

The comparison group growth expressed in standard scores, or expected growth, varies by grade, subject matter, test, testing cycle, and initial percentile. Supposedly, the procedures for calculating the RGI's take this into account by standardizing all scores. But, under some conditions expected growth will be quite small (near zero or even negative). These low indices of expected growth occur at the upper grade levels. They can result in extremely high (or low) measures of percent additional growth (or RGIs). As a result, the conventional wisdom that programs are "more effective" at the lower grade levels will often be reversed. But even more strikingly, RGIs can be undefined (where Expected Growth is zero) or so high that they are not credible (e.g., over 700%).

To illustrate, Table 1 presents example RGIs calculated using three different achievement tests, the California Achievement Test (CAT) Forms E & F, the Survey of Basic Skills (SBS) Forms P & Q, and the Metropolitan Achievement Test Version 6 (MAT6). Although the data are hypothetical, the results most likely understate the problem. We arbitrarily selected four grade levels for the analysis (grades 2, 6, 10 and 12). We then

identified national average NCE gains in reading (annual testing cycle) for 1985-86. We also identified the corresponding average pretest and posttest NCEs. Then, using the procedures outlined in Volume II of the Bilingual Education Evaluation System User's Guide (BEES), we calculated the RGIs corresponding to the NCE gains for the four selected grade levels. We calculated these RGIs for each of the three achievement tests just as if we had used that test and got results that matched the national averages. Tables 2, 3, and 4 include the NCE gain, the RGI, and most of the intermediate statistics such as the pretest and posttest standard scores corresponding to the 50th percentile "comparison group", the pretest and posttest standard scores corresponding to the project group's mean pretest and posttest NCE, the standardized pretest and posttest gaps, and the standardized growth of the national norm group. Figure 1 depicts the general gap-reduction model as described in the BEES documentation. Table headings may be referenced against Figure 1.

The results are striking. Given the same NCE gain and same pretest and posttest status but different tests, RGIs are highly variable. NCE gains convert to RGI losses, NCE losses convert to RGI gains, and depending on which tests standard scores are used RGIs for one grade level (grade 12) ranged from -92% to +11%.

These analyses show that project evaluation results will differ between NCE gains or RGIs. Yet, exactly the same set of test scores will have been used. One could argue that the same treatment effect (i.e. NCE gain) is harder to attain under different conditions and that is what is being reflected in the NCE/RGI discrepancies. But in the absence of a more internally valid design we may never know how to interpret the results. Clearly these data suggest the need for a technical investigation before requiring use of the gap-reduction calculations. They also point to a major problem for implementing the program improvement requirements of the new Chapter 1 law. Using different indices derived from the same pre and posttest data would result in different rankings for programs in terms of their need for improvement. What guidance should be given to practitioners seeking conclusions about program effectiveness? The logic of the RGI in relation to NCE gains has not been analyzed.

Conditions for Aggregation

In the BEES documentation, Tallmadge, Lam, and Gamel (1987) state that it is inappropriate to aggregate NCE gains for Bilingual programs because norms do not provide a valid comparison group. They recommend using RGIs instead. But with Chapter 1 programs serving students who are represented in national norms it has been considered acceptable to aggregate NCEs. Part of the rationale for aggregating NCEs is that they are an equal interval scale. Yet if RGIs transformed from NCEs yield different rank order than the NCEs both scales can not be equal interval by the same model. Thus, the paradox. How can results be validly aggregated to produce a state or federal report? Even if we can have it both ways, should RGIs of 700 be included in an aggregation? Where do we draw the line on what can be aggregated or meaningfully interpreted? These questions should be addressed through technical analysis.

Data Quality

Requiring a local project to compute RGIs means that in addition to conducting the Model A analysis, it must: (1) look up standard scores for four data points and standard score standard deviations for two occasions (pre and post), (2) compute both means and medians for the project group pretest and posttest distributions, (3) decide whether to use means or medians in the analysis, (4) compute a pooled standard deviation, which is then used to divide into the gap-reduction. Even with the use of a computer, to which many small projects will not have access, the opportunities for error in reporting are significant. The demand for training will be heavy and, given the interpretation problems outlined above, it does not appear that the cost benefit of the procedure could be justified. As an alternative, calculations could be done at the SEA level with TAC assistance. However, SEA-level calculations of RGIs would have its own set of data quality problems.

We have not done a thorough analysis of the BEES manuals and we do realize that they have been subject to careful review. Still, some further points may be made. The first is that the claims and requirements for the model seem inconsistent. Some of the inconsistency appears to relate to the three stages of gap-reduction analysis which are: (1) the simple gap-reduction, (2) the percent of gap-reduction, and (3) the relative growth index. For example, see the claim that gap-reduction can be used with nontest outcomes such as attendance rates, course grades and other such measures (see page 83 of Vol. 1). But where these measures do not assess growth, the analysis must stop short of calculating RGIs. Unfortunately, that limitation is not spelled out in the section on using gap-reduction for nontest data where it should be. Rather it appears in the section on assessing gap-reduction using test scores. Another example is the statement that NCEs must not be used for the gap-reduction model analysis. Why can't NCEs be used for stage 1 and stage 2 analysis?

One final comment related to this last area of concern. It seems out of context to advocate use of the correction for regression in the gap-reduction analysis. First, the model is presented as a way to determine whether a project has met its objective of closing some gap, not for validly assessing a treatment effect. The correction for regression is a method for isolating a treatment effect. Secondly, the regression correction does not work uniformly in practice. It overcorrects for regression due to selection. On this point, we have enclosed another paper from a previous TAC technical investigation (Gabriel, Estes and Dush, 1984).

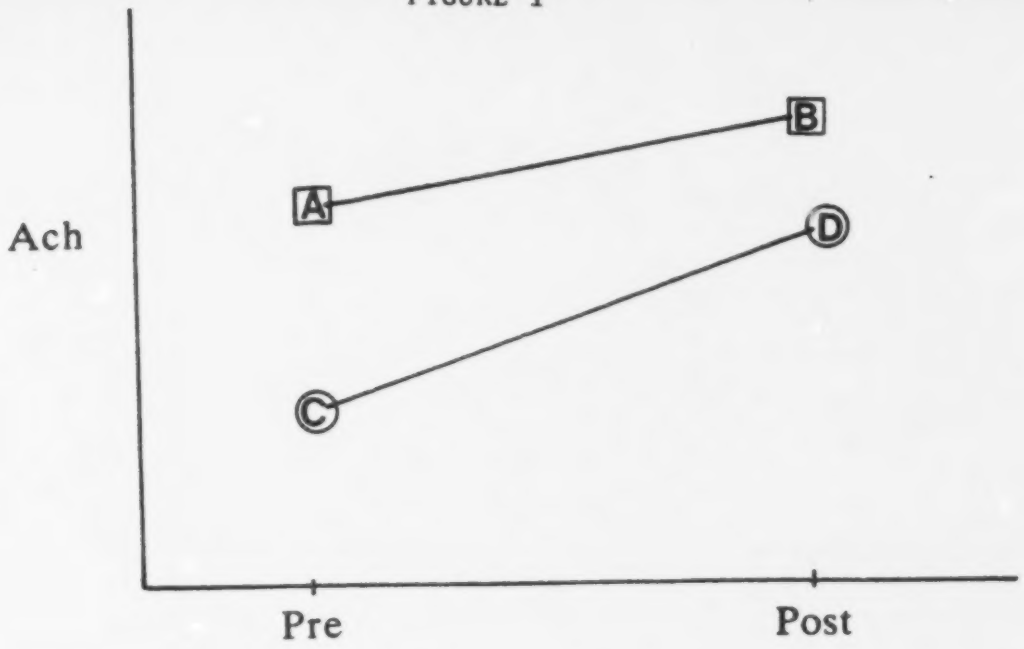
Recommendations

We hope that these comments can be used constructively. Our overall conclusion is that, while the model includes some features we have long believed in, there are some major problems with mandating the use of the gap-reduction model to extend TIERS and to aggregate results. On the other hand, technical investigations would help identify cases for which the gap-reduction model would augment local evaluations.

Recommendation #1: Do not require aggregating RGIs. Rather, encourage piloting the gap-reduction model as a means to augment local evaluations. In so doing attempt to encourage cases in which local comparison groups and nontest outcomes are used. TACs could help interested LEAs through a technical investigation.

Recommendation #2: Encourage LEAs to use annual testing and explicitly allow selection on the pretest at the local level. Then either leave the gains "uncorrected" acknowledging that they are measures of relative gain and not valid measures of treatment effect or adjust the results downward (1-2 NCEs) to better estimate treatment effects.

FIGURE 1



Pretest Gap = (A - C)

Posttest Gap = (B - D)

Gap Reduction = (A - C) - (B - D)

$$\text{Percent Gap Reduction} = \frac{(A - C) - (B - D)}{(A - C)} * 100$$

$$\text{Relative Growth Index} = \frac{(A - C) - (B - D)}{(B - A)} * 100$$

Gap Reduction Model

Stages of Analysis

TABLE 1

SAMPLE RGIs

GRADE	NCE GAIN	RGI		
		MAT-6 F & M	CAT E & F	SBS P & Q
2	1.1	- 1.7%	6.6%	10.8%
6	3.5	66.4%	48.6%	105.8%
10	1.3	73.9%	25.7%	30.8%
12	- .3	.0%	-91.8%	11.6%

TABLE 2

SAMPLE GAP REDUCTION ANALYSES

CALIFORNIA ACHIEVEMENT TEST: FORMS E & F
READING COMPREHENSION

GRADE	NCE GAIN	RGI	$(A-C)/sd_1$	$(B-D)/sd_2$	(B-A)/sd	sd1	sd2	Pooled sd	A	B	C	D
2	1.1	6.6%	.61	.54	1.15	96.3	65.3	82.3	545	640	486	605
6	3.5	48.6%	.82	.61	.43	43.9	34.4	39.4	722	739	686	718
10	1.3	25.7%	.91	.87	.19	22.0	20.8	21.4	765	769	746	751
12	-.3	-91.8%	.80	.85	.05	20.0	20.1	20.1	772	773	756	756

TABLE 3

SAMPLE GAP REDUCTION ANALYSES

SURVEY OF BASIC SKILLS: FORMS P & Q
READING COMPREHENSION

GRADE	NCE GAIN	RGI	$(A-C)/sd_1$	$(B-D)/sd_2$	(B-A)/sd	sd1	sd2	Pooled sd	A	B	C	D
2	1.1	10.8%	.60	.48	1.11	124.0	80.5	104.5	461	577	386	538
6	3.5	105.8%	.83	.57	.25	78.5	67.0	73.0	724	742	659	704
10	1.3	30.8%	.90	.80	.33	31.5	30.0	30.8	784	794	757	770
12	-.3	11.6%	.89	.86	.26	32.5	29.0	30.8	797	805	768	780

TABLE 4

SAMPLE GAP REDUCTION ANALYSES

METROPOLITAN ACHIEVEMENT TEST: FORMS F & M
READING COMPREHENSION

GRADE	NCE GAIN	RGI	$(A-C)/sd_1$	$(B-D)/sd_2$	$(B-A)/sd$	sd1	sd2	Pooled sd	A	B	C	D
2	1.1	- 1.7%	.61	.63	1.24	42.5	39.5	41.0	512	563	486	538
6	3.5	66.4%	.75	.56	.28	44.0	48.0	46.0	635	648	602	621
10	1.3	73.9%	.86	.80	.08	48.5	50.0	49.3	680	684	637	644
12	-.3	.0%	.85	.85	.12	50.5	50.5	50.5	696	702	653	659